简体中文 LIWC2024（SCLIWC2024）词典的修订与验证

崔雪婷[1,2] 陈思仪[1,2] 赵楠[1,2] 刘晓倩[1,2] 朱廷劭[1,2*]
[1] 中国科学院心理研究所 北京
[2] 中国科学院大学心理系 北京

摘要：近年来，字词分析取向的方法逐渐受到重视，特别是语言探索与字词计数(Linguistic Inquiry and Word Count, LIWC)工具，它的问世让许多心理学家对语言分析研究重新燃起热情。最新版本 LIWC-22 词典的修订新增了许多心理变量，在增加了 LIWC 工具的应用潜力的同时也使其更加完善。为进一步推动 LIWC 工具中文化的进程，我们对多个版本的中文 LIWC 词典进行汇总，修订形成了 SCLIWC2024，并对其效度进行了检验。研究一中，我们对照 LIWC-22 词典和 CLIWC2015 词典，以 SCLIWC 词典为基础，修订形成了 SCLIWC2024 词典。研究二中，我们进行了两项实验来检测 SCLIWC2024 在不同类型网络文本心理表达的有效性，并回答了如何更有效地使用 SCLIWC2024 来检测社交网络平台短文本的心理表达的重要问题。

关键词：LIWC，词典构建，SCLIWC2024，效度检验，网络文本分析

The Revision and Validation of the Simplified Chinese Linguistic Inquiry and Word Count Dictionary 2024(SCLIWC2024)

Cui Xueting[1,2] Chen Siyi[1,2] Zhao Nan[1,2] Liu Xiaoqian[1,2] Zhu Tingshao[1,2*]

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing

Abstract. In recent years, the Linguistic Inquiry and Word Count (LIWC) tool has garnered increasing attention, offering the promise of objective, automated, and transparent psychological text analysis. This resurgence has reignited enthusiasm among psychologists for language analysis research. The recent revision of the LIWC-22 dictionary has introduced numerous variables aimed at assessing various socio-psychological structures, thus expanding the application potential of the LIWC tool. To further promote the cultural adaptation of the LIWC tool, we have revised and validated the Simplified Chinese Linguistic Inquiry and Word Count Dictionary 2024 (SCLIWC2024) to better align with the features of LIWC-22. In Study One, building upon the SCLIWC dictionary, we revised SCLIWC2024 by comparing it with the LIWC-22 and CLIWC2015 dictionaries. In Study Two, we conducted two experiments to validate the efficacy of SCLIWC2024 in detecting different psychological semantics in online texts, addressing crucial questions regarding how to more effectively utilize SCLIWC2024 for detecting the psychological semantics of short texts on social networking platforms.

Keywords：LIWC, revision, Simplified Chinese LIWC, validity verification, web text analysis

引言

语言是人们将内心思想和情感转化为他人可理解的形式中最常见、最可靠的方式(Tausczik & Pennebaker, 2010)。从历史上看，语言心理学的研究虽然从表达性写作研究演变而来(例如，Gottschalk & Gleser, 1969)，但已经扩展到了社会科学中的各种主题。与此同时，随着第一个心理学通用计算机文本分析程序 General Inquirer(Stone et al., 1966)的兴起衰落，字词分析取向方法逐渐受到重视，人们已经认识到通过计算不同含义和词性的词语的表达频率可以发现文本的心理意义(Pennebaker et al., 2014; Kacewicz et al., 2014)。

近年来，数字革命导致大规模语言数据的可用性增加(Boyd & Pennebaker, 2017)，语言心理学研究也已经扩展到更现代的交流环境中，例如在线社交媒体(Caton et al., 2015; Stanton et al., 2017)。一些真正对用户友好的文本分析软件一一面世、层出不穷，特别是由 Pennebaker 的研究团队发展出来的语言特性计算机分析工具——LIWC(最新版本为 LIWC-22，Boyd et al., 2022)，近年来尤其受到瞩目。事实上，研究者目前对言语行为和心理之间关系的大部分理解都直接或间接地建立在 LIWC 单词计数方法上，它使客观、自动化和透明的心理文本分析的前景成为了现实(Boyd & Schwartz, 2021)。

LIWC 研究工具的核心在于扫描文本，并从预定义的类别中统计单词的频率。LIWC 词典定义了各个来源于心理学理论的类别名称及各类别中所包含的字词列表，因此，每个类别的相对频率通常反映了一个人对每个领域的相对关注程度。最新修订的 LIWC-22 词典(Boyd et al., 2022)由 12000 多个单词、词干、短语和精选的表情符号组成，包括情感、认知、文化、生活方式等 117 个在心理学或语言学上有意义的类别以及 4 个未公开的复合式指标，旨在评估各种社会心理结构。

这套研究工具的重要优势在于，它分为程序主体和词典两个部分，允许用户扩展词典或添加新的类别，在搭载各国语言词典后，可用以分析不同语言的文本(例如，Meier et al., 2018; Date et al., 2024)。对于中文词典的开发，黄金兰等人(2012, 2020)已针对 LIWC2007 和 LIWC2015 分别进行了相应版本词典的中文化翻译修订工作，形成了 CLIWC2007 词典和 CLIWC2015 词典。Gao 等人(2013)也根据 LIWC 词典和 CLIWC2007 词典建立了专门的简体中文 LIWC 词典(SCLIWC)，并将微博高频词添加到了词典中。CLIWC 词典的建立为 LIWC 中文

化奠定了良好的基础，使计算机文本分析方法应用于繁体中文文本分析的相关研究成为可能，SCLIWC 词典则在 LIWC 研究工具和简体中文之间架起了桥梁。

在过去的 20 多年里，LIWC 以及其他语言版本的词典已经越来越普遍地被应用于不同的研究领域中，特别是更新后的 LIWC-22 版词典又引入了许多新的类别。LIWC 词类所描述的语言特征反映了人们的注意焦点(Brockmeyer et al., 2015)、认知过程(Konopasky et al., 2020)、情感情绪(Sun et al., 2020)、社会行为(Samson et al., 2023)、状态与动机(Cherrstrom, 2023)、生活方式(Varma et al., 2021)、心理健康(O’Dea et al., 2017)、诚实与欺骗(Markowitz & Griffin, 2020)、思维方式(Markowitz, 2023)、个体差异(Ulrichová & Havigerová, 2023)等心理与行为过程。在计算机领域，LIWC 也被用作语言模型，用于预测人格特质(Hall & Caton, 2017)、沟通行为(Biggiogera et al., 2021)、心理健康(Lyu et al., 2023)、身体健康(Ziemer & Korkmaz, 2017)以及政治竞选结果(Jordan et al., 2018)等。LIWC 在各个领域的研究和数以千计的引用次数都足以证明其在语言分析研究中的重要地位。

在 LIWC 词典中文化的工作上，虽然已有黄金兰等人(2012, 2020)建立的 CLIWC2007 和 CLIWC2015 词典、Gao 等人(2013)建立的 SCLIWC 词典，但随着 LIWC-22 英文版词典的修订，Boyd 等人(2022)不仅建立了多个全新词类，增加了 LIWC 的应用潜力，同时修订了原有类别涵盖的单词列表，使其更加完善。因此，我们的工作重点是借助词向量大模型，综合心理学和计算机科学方法，根据 LIWC-22 英文词典和 CLIWC2015 中文词典对 SCLIWC2024 词典进行更新和效度检验，以符合 LIWC-22 的特性。

研究 1 SCLIWC2024 的建立

本研究的目的是基于 LIWC-22 词典和 CLIWC2015 词典修订 SCLIWC2024 词典，下面我们将详细说明 SCLIWC2024 词典的修订过程。

（一）添加新增词

第一阶段的修订旨在依据 CLIWC2015 修订原则对 SCLIWC2024 词典进行新增词类和新增词添加。

CLIWC2015 词典共 79 个词类，其中 61 个词类与旧版 SCLIWC 词典相同，18 个词类为新增词类。参考 CLIWC2015 词典修订手册，并两两比对 CLIWC2007、CLIWC2015、旧版 SCLIWC 词典的词类和词，我们对 SCLIWC2024 词典做出以

下修订。

第一，遵循 CLIWC2015 添加新词类的原则，我们在 SCLIWC2024 中直接增加了驱力词(Drives)、联系词(affiliation)、权力词(power)、酬赏词(reward)、风险词(risk)、女性词(female)、男性词(male)、非正式词(informal)、网络用词(netspeak)、、比较词(compare)、疑问词(interrog)和差异词(differ)共 12 个新词类，并直接使用 CLIWC2015(简体)中这 12 类的词。需要注意的是，非正式词(informal)在第二阶段与 LIWC-22 融合过程中更名为会话词(Conversation)。

第二，对于时态标定词类，我们将旧版 SCLIWC 词典中 PastM、PresentM、FutureM 词类的词分别与 CLIWC2015 中的 focuspast、focuspresent、focusfuture 合并去重，形成三个新的词类：过去(focuspast)、现在(focuspresent)、未来(focusfuture)。

第三，CLIWC2015 将 CLIWC2007 的多功能词类(MultiFun)重新分类到了语助词类 Particle(Interjunction)，并进一步区分为情态语助词(Modal_pa)与一般语助词(General_pa)两个子词类。针对这一点，我们直接使用 CLIWC2015 中的 Particle、Modal_pa、General_pa 三个词类，替换掉了 SCLIWC2024 中的 Interjunction 与 MultiFun。

第四，对于各个词类下的新增词，通过比对简体和繁体 CLIWC2015，我们发现简体版和少部分繁体版翻译后的新增词符合大陆普通话习惯且覆盖较全，因此我们将 CLIWC2015 剩余 61 个词类中的新增词合并去重后直接添加进了 SCLIWC2024 对应的词类。

经过第一阶段修订后的 SCLIWC2024 词典包含 84 个词类，共计 14229 个词。需要注意的是，针对 CLIWC2015 删除的词类以及拆分合并的词类，我们在 SCLIWC2024 中仍予以保留，只增加 CLIWC2015 的新增词类。

（二）融合 LIWC-22 词典

第二阶段的修订旨在依据 LIWC-22 词典修订原则对 SCLIWC2024 词典进行新词类(词)增加和旧词类更新。

相比于以往的任何版本，LIWC-22 的每个方面都经历了重大更新。在词典整体结构上，LIWC-22 将类别划分为“Basic”和“Expanded”超类别。在概念上，Basic 类包含了早期 LIWC 版本的大多数维度，Expanded 类则包括了更新后

的传统 LIWC 类别以及其它新增类别(详见 Boyd et al., 2022)。根据 LIWC-22 的修订过程，我们遵循三个原则对 SCLIWC2024 进行修订。一是将 LIWC-22 的新增词类、更新的旧词类(共计 32 个)直接翻译，筛选后加入或修改 SCLIWC2024，并根据 LIWC-22 词典修改 SCLIWC2024 词典结构；二是逐类对比 LIWC-22 和 LIWC2015，将旧词类的新增词加入 SCLIWC2024；三是保留 LIWC-22 删除的词类：比较词(compare)、疑问词(interrog)、相对词(Relative)。

在翻译过程中，考虑到中英文在语言特性上的差异，我们删除了形容词(adj)、冠词(article)和 LIWC-22 新增的限定词(det)三个词类。此外，LIWC-22 的四个总结变量都来源于 Pennebaker 团队之前发表的研究结果(Boyd et al., 2022)，但并未透明公开，因此也不予以考虑。最终，我们总共整理翻译了 99 个 LIWC-22 词类。

在完成对 LIWC-22 词典的翻译后，我们计算了每个词的相似度以辅助进行词类适合度评估。具体来说，对每一词类，我们选取了其英文词类名称或特征作为标准词，例如成就词(achieve)的标准词为 achieve，消极情绪(emo_neg)的标准词为 negative。然后我们对词类中每个单词的释义词向量与标准词词向量分别计算余弦相似度并进行高低排序。

在词条相似度排序的基础上，我们训练了两名研究生对每个词类中所有的翻译词条进行了词类适合度的人工评判。具体来说，我们将词条平均分为两组，每组各由一名研究生进行评判，完成后再互换词条进行评判。对每个词条项目，均会得到两个判断反应，若判断反应一致(全部适合/全部不适合)，则予以保留/删除。然后，再由两位工作人员对判断反应不一致的词条项目进行判断，将反应一致的予以保留/删除。

以往 LIWC 研究显示，过去常被忽视的功能词类及其组合词类，具有相当不错的心理特性表征功能(Tausczik & Pennebaker, 2010)。因此，我们还重新修订了 SCLIWC2024 的功能词(function)。考虑到不同语言的使用差异，我们没有直接使用 LIWC-22 翻译得到的功能词，而是参考相关简体中文词典书籍(岑玉珍, 2013; 方清明, 2017; 刘月华 等, 2019; 齐沪扬, 2011; 中国社会科学院语言研究所词典编辑室, 2016)，对照汉语语法类别使用介绍与词典说明，综合统整出分属各功能词类的词语。

接下来，参考 CLIWC2007 的建立(黄金兰 等, 2012)，我们完成了

SCLIWC2024 的单一词确定。首先，我们对现有中文词条项目进行分词处理，以此作为单一词判断依据。然后，根据词条所属类别以及《现代汉语词典（第 7 版）》(中国社会科学院语言研究所词典编辑室, 2016)对被断成两个词及以上的翻译词进行保留、删除或修改。

对于词典中的词干，LIWC 利用词干加星号的方式来侦测同词干的词群。CLIWC 词典(黄金兰 等, 2012, 2020)则以词首字为词干(或称词素)来标示相对应的概念，例如，用悲*来侦测悲伤、悲痛、悲哀、悲恸、悲苦等悲伤词。参考 CLIWC 词干建立的方法，我们对 SCLIWC2024 补充了新的词干。我们首先根据 SCLIWC2024 现有词干侦测了 LIWC-22 新增词并予以删除，然后对剩余每个词类的新增词建立词干。我们参考了《现代汉语词典（第 7 版）》(中国社会科学院语言研究所词典编辑室, 2016)，若“某词*”下所有的词均可归为同一类(几类)，则建立该词干。需要注意的是，CLIWC 建立的词干只包括首词加“*”的格式，但查阅《现代汉语词典 第七版》后，我们认为也可以建立“*”加末词格式的词干，如“*慕”(爱慕、渴慕、倾慕)。在词干确立过程中，我们再次对词条的词类适合度进行了确认。

最后，我们将按上述步骤得到的新增词类、词、词干进行了整理。首先，我们将上述通过词类适合度检验的新增词类(词)和通过词性验证的功能词(function)直接合并去重加入 SCLIWC2024 词典。然后我们对 Linguistic、function 词类进行了最终的词类结构检查，将所有下级词类的词加入上级词类，同时整理出上级词类特有的词再进行词性确认。最终，经过第二阶段修订后的 SCLIWC2024 词典包含 116 个词类，共计 18105 个词。

（三）大语言模型扩充词典

第三阶段的修订旨在根据腾讯 AI Lab 大模型对 SCLIWC2024 词典进行扩充。

腾讯 AI Lab 此次公开的中文词向量数据包含 800 多万中文词汇，其中每个词对应一个 200 维的向量。相比现有的中文词向量数据，腾讯 AI Lab 的中文词向量着重提升了覆盖率、新鲜度和准确性三个方面，相比已有各类中文词向量大大改善了其质量和可用性(Song et al., 2018)。

由于 Linguistic 维度下部分词类是根据词性建立的，不能根据近义词扩展，因此，我们排除了部分功能词，仅拓展了 94 个词类。首先，我们对每个词类的

所有词进行了近义词拓展，然后对所有拓展词进行基于 LTP 分词(Che et al., 2010)的单一词判断，最后对拓展后的近义词进行词类适合度判断。最终通过合并去重加入 SCLIWC2024 的新词只有 2883 个。

（四）陈旧词处理和词类结构检查

第四阶段的修订旨在处理 SCLIWC2024 词典中的陈旧词并进行词类结构检验。

对于陈旧词，SCLIWC2024 词典中有些词已经很少被人们使用，因此我们谨慎地去掉了某些词。由于这些词大多为名词，且多集中在扩展类别(Expanded)，因此我们选择生活方式(lifestyle)下的词类和网络用词(netspeak)，根据三条标准进行陈旧词筛选：一是随时间变化词义发生改变、不再适合该词类的词；二是过时、不再流行的词；三是使用场景单一的词。在去除陈旧词后，我们对 SCLIWC2024 的词类结构进行了最终检查，确保上级词全部包含下级词，并对上级特有词进行了下级词类适合度判断。

最终，经过上述四个阶段修订的 SCLIWC2024 词典共包含 116 个词类，其中语言维度包含 27 个词类，心理过程包含 41 个词类，扩展类别(Expanded)包括 48 个词类，共计 19952 个词。表 1 列出了各个词类的名称和范例。

表 1 SCLIWC2024 的语言维度和词类

变量名称	Category	简称	范例	词数
语言维度	**Linguistic Dimensions**	Linguistic		2616
功能词	Total function words	function	这、它、还有	2614
代名词	Total pronouns	pronoun	我、他们、自己	145
特定人称代名词	Personal pronouns	ppron	我、他、你	61
第一人称单数代名词	1st person singular	i	我、本人、自己	17
第一人称复数代名词	1st person plural	we	我们、咱、咱们	8
第二人称代名词	2nd person	you	你、您、你的	12
第二人称复数代名词*	YouPL	YouPL	你们、汝等、尔等	7
第三人称单数代名词	3rd person singular	shehe	他、她、他的	12
第三人称复数代名词	3rd person plural	they	他们、她们、人家	9
非特定人称代名词	Impersonal pronouns	ipron	其他、那些、彼此	90
数字	Numbers	number	一、两、三	130
介系词	Prepositions	prep	在、和、除了	117
助动词	Auxiliary verbs	auxverb	要、可以、应该	82
副词	Adverbs	adverb	都、无论、虽然	700
连接词	Conjunctions	conj	一旦、并且、虽然	172
否定词	Negations	negate	不、没有、没	61
动词	Common verbs	verb	分享、说、到	818

续表 1

变量名称	Category	简称	范例	词数
量词	Quantities	quantity	稀少、一些、不等	233
比较词	Comparisons	compare	先前、相异、好比	188
疑问词	Interrogatives	interrog	什么、还是、怎么	32
后置词*	PrepEnd	PrepEnd	期间、以后、为止	52
特指定词*	SpecArt	SpecArt	这、什么、那	27
语助词*	Particle	Particle	的、了、着	113
情态语助词*	Modal_pa	Modal_pa	呢、吧、吗	101
一般语助词*	General_pa	General_pa	的、了、着	12
数量单位词*	QuanUnit	QuanUnit	对、个、里	154
心理过程	**Psychological Processes**			
驱力词	Drives	Drives	主宰、让、工作	2472
联系词	Affiliation	affiliation	我们、社会、孩子	594
成就词	Achievement	achieve	优秀、冠军、实现	581
权力词	Power	power	上司、让、允许	1498
认知词	Cognition	Cognition	决不、考虑、熟记	2500
绝对主义思维	All-or-none	allnone	都、不、没有	48
认知历程词	Cognitive processes	cogproc	意料、明白、因为	2374
洞察词	Insight	insight	发现、了解、觉得	694
因果词	Causation	cause	为什么、因为、结果	259
差距词	Discrepancy	discrep	必须、欠缺、应该	163
犹豫词	Tentative	tentat	大概、也许、猜想	311
确切词	Certitude	certitude	一定、绝对、就是	351
限制词	Inhibition	Inhibition	废止、不准、规则	210
包含词	Inclusive	Inclusive	包括、附近、添加	157
排除词	Exclusive	Exclusive	取消、但是、除外	75
差异词	Differentiation	differ	例外、反而、差别	201
记忆词	Memory	memory	记忆、忘记、失忆	42
情感历程词	Affect	Affect	喜欢、愤怒、悲伤	3814
正向感情	Positive tone	tone_pos	欢乐、爱、得意	1524
负向感情	Negative tone	tone_neg	自负、失败、怠慢	1897
情绪词	Emotion	emotion	欢呼、啜泣、崩溃	1204
积极情绪	Positive emotion	emo_pos	感动、高兴、欢喜	280
消极情绪	Negative emotion	emo_neg	紧张、焦躁、疲倦	919
焦虑词	Anxiety	emo_anx	不安、担心、焦虑	208
生气词	Anger	emo_anger	愤怒、讨厌、可恶	440
悲伤词	Sadness	emo_sad	伤心、哭、凄惨	273
脏话	Swear words	swear	智障、混蛋、靠	193
社会历程词	Social processes	Social	公认、大家、揭露	3671
社会行为	Social behavior	socbehav	支持、关系、合作	2305
亲社会行为	Prosocial behavior	prosocial	救济、资助、帮	298
礼貌词	Politeness	polite	道歉、谦恭、请	170
冲突词	Interpersonal conflict	conflict	嫌隙、争夺、击败	363
道德词	Moralization	moral	无能、自负、诚实	523

续表 1

续表 1

变量名称	Category	简称	范例	词数
沟通词	Communication	comm	说服、报告、答辩	570
社会性参照	Social referents	socrefs	我、伙伴、人	1419
家庭词	Family	family	家人、妈妈、结婚	292
朋友词	Friends	friend	朋友、同事、交往	154
女性参照	Female references	female	女人、母亲、女友	263
男性参照	Male references	male	男人、父亲、丈夫	210
人类词	Humans	humans	人民、成员、群众	329
相对词	Relativity	Relative	中间、加快、以往	2450
扩展类别	**Expanded Dictionary**			
文化词	Culture	Culture	中国、社会、国家	1013
政治词	Politics	politic	安全、治理、政策	411
种族词	Ethnicity	ethnicity	中国、美国、欧洲	379
科技词	Technology	tech	技术、汽车、网络	227
生活方式	Lifestyle	lifestyle	发展、建设、市场	3778
休闲词	Leisure	leisure	旅游、消费、体育	1281
房屋词	Home	home	家、居民、房子	252
工作词	Work	work	工作、企业、合作	1558
金钱词	Money	money	经济、钱、价值	575
宗教词	Religion	relig	灵魂、主、神	416
生理历程词	Physical	physical	身体、活力、头	2874
身体词	Body	body	心、血液、大脑	573
健康词	Health	health	健康、医疗、药品	943
疾病词	Illness	illness	医院、患者、疾病	502
健康保养	Wellness	wellness	生活、幸福、运动	151
心理健康	Mental health	mental	心理、发作、诊断	113
药物词	Substances	substances	酒、抽烟、吸毒	140
性词	Sexual	sexual	做爱、吻、抱	257
摄食词	Food	food	吃、粮食、水	795
死亡词	Death	death	死、烈士、墓	288
状态	States			
需要词	Need	need	需要、应该、基本	77
欲望词	Want	want	想、期望、追求	37
获得词	Acquire	acquire	取得、拥有、达到	185
缺少词	Lack	lack	没有、少、不够	82
满足词	Fulfilled	fulfill	多、超过、丰富	50
疲倦词	Fatigue	fatigue	累、叹气、无聊	68
动机	Motives			
酬赏词	Reward	reward	利润、报酬、奖金	142
风险词	Risk	risk	安全、风险、危机	224
好奇心	Curiosity	curiosity	发现、创造、挑战	74
吸引词	Allure	allure	喜欢、享受、品牌	307
感知历程词	Perception	Perception	刺耳、柔软、可口	3924
注意词	Attention	attention	关注、注意、吸引	141

变量名称	Category	简称	范例	词数
移动词	Motion	motion	下降、通过、行进	1059
空间词	Space	space	方位、中央、边界	1235
视觉词	Visual	visual	看、明显、亮	600
听觉词	Auditory	auditory	说、告诉、响	501
感觉词	Feeling	feeling	温暖、热、辣	416
时态标定词*	TenseM	TenseM	时候、持续、之前	1018
延续*	ProgM	ProgM	已经、近日、至今	13
时间词	Time	time	一年、世纪、下旬	883
过去	Past focus	focuspast	当时、过去、昨天	83
现在	Present focus	focuspresent	现在、今年、目前	46
未来	Future focus	focusfuture	未来、然后、明天	122
会话词	Conversational	Conversation	粉丝、微博、好的	913
网络用词	Netspeak	netspeak	靠、梗、热搜	780
应和词	Assent	assent	好、可以、好的	95
停顿赘词	Nonfluencies	nonflu	这、啊、嗯	24
填充赘词	Fillers	filler	就是、对于、所以	42

注：*标识的词类代表中文词典特有的类别，共 9 个。

研究 2 SCLIWC2024 的效度检验

LIWC 的有效性一直都是一个至关重要的问题。从定义上讲，词典的每个词类都是有效的，但我们仍需回答一个问题：当使用某类别的词汇时，是否意味着人们确实按照该类别定义的方式表达了含义？这个问题涉及个体心理过程如何反映在语言使用中，以及词类作为特征集的有效性。Pennebaker 和 Francis(1996)对 LIWC 的效度进行了最早的测试，通过计算人类和 LIWC 对写作内容评分的相关性，成功证明了 LIWC 在测量积极和消极情绪、认知策略等方面是有效的。人类评分和 LIWC 评分的一致性为 LIWC 的效度提供了支持。

自 LIWC 问世以来，尽管数千项研究和数百个独立实验已经证实了其各个类别的有效性(Boyd & Schwartz, 2021)，但当 LIWC 被翻译成其他语言时，我们仍需要提供该语言下的有效性数据。此外，任何文本分析系统的评估都需要抽取大量不同主题的文本样本，以广泛验证其有效性。在 LIWC-22 开发过程中，Boyd 等人(2022)建立了 Test Kitchen 语料库，它包含了 15 个从正式到非正式的主题，几乎涵盖了所有的网络文本形式，广泛地代表了人们在日常生活中使用单词的方式，并且已被用于验证 LIWC-22 的效度。

因此，为了满足处理简体中文文本，特别是网络文本的需要，我们复制了

Zhao 等人(2016)对 SCLIWC 的效度检验，旨在回答两个问题：SCLIWC2024 检测不同网络文本心理表达的有效性如何？以及如何更有效地使用 SCLIWC2024 来检测社交网络平台短文本的心理表达。

实验 1

本实验旨在评估 SCLIWC2024 对网络文本心理表达识别的有效性。我们通过 SCLIWC2024 和人工评分者分别处理三种不同的网络文本(短文本、长文本、传统网络文本)，以评估 SCLIWC2024 在不同文本状态下检测心理表达的效度。同时，我们在不同时间跨度的微博状态下对 SCLIWC2024 不同评分方法的效度进行了评估和比较。

方法

参与者和实验材料。我们选取了在中国非常受欢迎的社交媒体平台——新浪微博以及主流新闻网站——人民日报和南方周末。

(1) 短文本

我们随机抽取了 120 名微博用户(男性 60 名，女性 60 名)在 2021 年 12 月 1 日至 2021 年 12 月 31 日的微博状态，这些用户是从我们活跃的微博用户池中随机抽取的(Li et al., 2014)，并且符合以下要求：

A. 在 2021 年 12 月 1 日至 2021 年 12 月 31 日期间，每个用户发布了 90-110 条有效状态。

B. 在 2021 年 12 月 1 日至 2021 年 12 月 7 日(第一周)期间，每个用户发布了 20-26 条有效状态。

C. 在 2021 年 12 月 1 日，每个用户发布了 3-4 条有效状态。

此处的有效状态是指在删除链接、转发内容(“//@用户名：”或通过新浪微博 API 在下载数据对象的“retweeted_status”字段中进行标记）、提及(“@用户名”)和表情图标后字数大于 0 的状态。我们关注的是 SCLIWC2024 处理文本的有效性，因此删除了链接、表情图标、图片、音频或视频等超出 SCLIWC2024 处理范围的内容。此外，提及他人的用户名、转发内容通常夹杂着广告、新闻等非个人表达，也予以删除。预处理后样本的平均文本长度为 34.85，范围从 2 到 234，男性用户(98)的状态数与女性用户(99)基本相等，但文本长度男性(44.60)略高于女性(31.72)。

(2) 长文本

我们在微博平台的长文本板块中随机选取了 120 名用户(男性 60 名，女性 60 名)的 120 篇长文本(M=1196.0, SD=518.95)，内容涉及作者的经历、思考和感受。

(3) 传统网络文本

我们选取了 2022-2024 年中国主流媒体网站(人民日报、南方周末)发布的 120 条新闻评论(M=1446.37, SD=770.36)，内容涉及时政、经济、军事、文化、社会热点等。

评定词类的选取。我们评估的重点是那些基于明确界定的心理过程构建的词类，而不是那些仅仅是一类词集合的词类(例如人称代词)。我们选取了 25 个 SCLIWC2024 词类进行评估，包括认知历程(allnone、insight、cause、tentat)、情感历程(emo_anger、emo_anx、emo_sad、emotion、tone_neg、tone_pos)、社会历程(conflict、moral、polite、prosocial)、生理历程(health)、生活方式(work)、驱力词(achieve、affiliation、power)、动机(allure、curiosity、risk)和状态(acquire、fatigue、fulfill)。以往研究中，认知历程、情感历程、社会历程和生理历程已经被用于预测个体的人格特质、思考过程、身心健康、人际关系及其它心理与行为指标，并展现了显著的预测效果(Tausczik & Pennebaker, 2010)。LIWC-22 的问世后，部分研究者也开始用 LIWC-22 中的新增词类来研究其他心理过程，例如绝对主义思维(Yahya & Abdul Rahim, 2023)、亲社会行为、冲突和道德等社会行为(Haupt et al., 2023; Samson et al., 2023)、动机和状态类别(Cherrstrom et al., 2023)。因此，我们增加了一些近年来研究者重点关注的 LIWC-22 新增类别。

人工评分。评分者需要根据每个词类的定义判断给定文本表现词类含义的程度。这些类别的定义参考了 LIWC-22 的构建手册(Boyd et al., 2022)。我们的评分者被要求评估整个文本内容的含义，而不是检测某些单词。例如，如果文本内容明显的表达了作者非黑即白的思考方式，那么不管文本中是否出现绝对主义思维词，它都会被归为绝对主义思维(allnone)，我们也排除了那些仅仅为了表达某种偏好而使用情感词汇的描述，例如“很喜欢这部电视剧”，以识别那些真正的情感表达。

我们训练了三名研究生，在不知道 SCLIWC2024 评分结果情况下对所有文本进行独立评分。对每个微博用户，评分者依次判断其第一天/第一周/整个月的

微博状态表达每个类别含义的程度(7 分制)。对每个长文本和新闻评论，评分者对整个文章也做出类似的判断。我们以 3 位评分者的评分均值作为最终人类评分分数，并使用 Cronbach alpha 对 3 个评分者的信度进行了测量，结果范围在 0.72(微博_一周_prosocial)和 0.99(传统网络文本_tone_pos)之间，表明评分者在三种文本类型中的评分一致性都是可接受的。

SCLIWC2024 评分。在对文本进行 LTP 分词后，我们使用 SCLIWC2024 对文本进行了两种分析。一是进行词数计分，即直接使用 SCLIWC2024 对微博短文本集(一个月)、长文本集和传统网络文本集进行分析，得到每个类别的单词百分比分数。二是进行状态数计分，对于单个微博状态，若其中出现一个或多个 SCLIWC2024 类别的单词，则将其标记为该类别，然后计算每个微博用户在每个词类下的状态数。

结果与讨论

作为使用 LIWC 工具分析文本的常用方法，我们首先对每个文本集计算了每个词类的单词百分比，并与人工评分进行比较。表 2 展示了三种文本下的的单词百分比。25 个词类根据 SCLIWC2024 词典结构被分为 8 个维度，大多数词类的单词百分比与 Boyd 等人(2022)对不同文本的分析结果相似。但在认知历程维度，绝对主义思维(allnone)、因果词(cause)、洞察词(insight)略高于 Boyd 等人(2022)的结果，尤其是在长文本中。在情感历程维度，正向感情(tone_pos)和负向感情(tone_neg)在三种不同文本中的单词百分比都略高于 Boyd 等人(2022)的结果。生活方式维度的工作词(work)和驱力词维度的三个词类的单词百分比都高于 pennebaker 等人的结果，尤其是在传统网络文本中。而动机维度中的吸引词(allure)的单词百分比要低于 pennebaker 等人的结果。

三种不同的文本类型的单词百分比也呈现出文本类型特征的差异。传统网络文本以客观理性的方式表达观点，使用较多的因果词(insight)。在内容上，微博短文本和长文本更具个性化表达，并且常用绝对化的语言来夸大自我感受，例如“每天都是丢人玩意”，所以使用较多的绝对主义思维词(allnone)。而传统网络文本更关注经济、政治等公共话题，因此传统网络文本使用更多的工作词(work)、成就词(achieve)、权力词(power)。

表 2 SCLIWC2024 词典在微博短文本、长文本和传统网络文本中检测到的单词百分比

	微博(月)	长文本	传统网络文本
认知历程			
allnone	**3.08(1.43)**	**4.39(1.32)**	1.53(0.92)
cause	1.39(0.74)	**2.29(1.07)**	**3.10(1.33)**
insight	2.73(1.27)	**4.78(1.85)**	**4.40(1.85)**
tentat	1.93(1.02)	3.31(1.45)	1.15(0.83)
情感历程			
emo_anger	0.46(0.38)	0.70(0.82)	0.29(0.54)
emo_anx	0.29(0.24)	0.49(0.56)	0.21(0.33)
emo_sad	0.44(0.41)	0.60(0.60)	0.12(0.18)
emotion	3.91(2.58)	2.99(1.62)	1.78(1.00)
tone_neg	**3.04(1.65)**	**4.42(2.05)**	**2.24(2.02)**
tone_pos	**8.42(3.79)**	**6.08(2.43)**	**8.88(3.46)**
社会历程			
conflict	0.29(0.31)	0.50(0.51)	0.37(0.75)
moral	1.82(1.13)	1.57(1.10)	1.38(1.30)
polite	1.72(1.13)	0.98(0.69)	0.78(0.72)
prosocial	1.29(0.98)	1.05(0.89)	1.62(1.37)
生理历程			
health	1.13(0.92)	1.39(1.38)	1.94(3.03)
生活方式			
work	**4.46(3.61)**	**4.50(2.91)**	**12.96(5.65)**
驱力词			
achieve	**2.62(1.34)**	**3.27(1.85)**	**6.02(2.96)**
affiliation	**3.25(2.35)**	**3.42(2.27)**	**3.23(2.35)**
power	**3.81(1.32)**	**5.13(2.08)**	**7.75(3.21)**
动机			
allure	**4.32(1.52)**	**4.59(1.39)**	**3.54(1.38)**
curiosity	0.19(0.18)	0.28(0.29)	0.40(0.37)
risk	0.68(0.45)	1.28(0.81)	1.10(1.29)
状态			
acquire	1.80(0.80)	2.60(0.94)	1.72(0.89)
fatigue	0.11(0.16)	0.15(0.26)	0.02(0.06)
fulfill	0.33(0.24)	0.47(0.36)	0.82(0.53)

注：微博短文本的结果是根据研究中收集到的每个微博用户的所有有效微博状态来计算的。加粗的数据表明 SCLIWC2024 单词百分比要高于/低于 LIWC-22。

为了检验 SCLIWC2024 在检测微博短文本、长文本和传统网络文本心理表达的有效性，我们对 SCLIWC2024 得分和人工评分进行了 Pearson 相关分析(表 3)。当使用词数计分作为 SCLIWC2024 评分方法时，除了动机和状态类别外，大部分词类在长文本和传统网络文本的相关性高于微博短文本(月)，且有更多的显著相关性。对于认知历程类别，长文本和传统网络文本中大部分词类的

SCLIWC2024 得分与人类评分之间的相关性显著且达到中等水平。对于情感历程，SCLIWC2024 得分与人类对三种文本类型的评分之间的相关性非常密切。对于社会历程，除了“道德词(moral)”外，其他三个词类在传统网络文本上显著相关，且高于在微博短文本和长文本中的相关性。对于生理历程类别，传统网络文本达到了较高的显著相关，而在生活方式类别，长文本有着中等显著相关性。对于驱力词，除“成就词(achieve)”在长文本不显著外，其它词类在长文本和传统网络文本中均有中等显著相关性。对于动机和状态类别，相关性普遍较低，且较少词类显著相关，只有“风险词(risk)”在传统网络文本上有高显著相关、“疲倦词(fatigue)”在长文本有中等显著相关。

目前的研究结果表明，SCLIWC2024 词数计分在检测微博短文本、长文本和传统网络文本心理表达方面是有效的。所选类别的单词百分比以及在不同文本之间的差异与 Boyd 等人(2022)和 Zhao 等人(2016)的结果相似，证实了 SCLIWC2024 的结构效度。此外，从表 3 可以看出，不同心理语义类别的有效性不同，这也与 Zhao 等人(2016)的研究一致，反映出不同心理语义类别在语言中不同的映射方式。

表 3 SCLIWC2024 词数计分与人工评分在三种文本之间的相关性

	微博(月)	长文本	传统网络文本
认知历程			
allnone	0.125	0.323**	0.151
cause	0.159	0.431**	0.272**
insight	0.127	0.511**	0.234*
tentat	0.154	0.189*	0.407**
情感历程			
emo_anger	0.236**	0.289**	0.705**
emo_anx	0.144	0.398**	0.546**
emo_sad	0.231*	0.644**	0.582**
emotion	0.223*	0.511**	0.383**
tone_neg	0.110	0.600**	0.586**
tone_pos	0.213*	0.395**	0.142
社会历程			
conflict	0.019	0.548**	0.593**
moral	0.103	0.394**	0.101
polite	0.113	0.115	0.372**
prosocial	0.086	0.09	0.342**
续表 3			
生理历程			
health	0.094	0.584**	0.838**
生活方式			
work	0.205*	0.560**	0.282**

	微博(月)	长文本	传统网络文本
驱力词			
achieve	0.017	0.11	**0.271****
affiliation	0.094	**0.462****	**0.304****
power	0.108	**0.476****	**0.370****
动机			
allure	0.127	0.126	0.127
curiosity	0.178	0.142	**0.234***
risk	0.178	**0.296****	**0.720****
状态			
acquire	0.056	0.162	0.117
fatigue	0.099	**0.359****	0.05
fulfill	0.018	0.128	0.15

注：加粗的数据表明 SCLIWC2024 分数与人类评分显著相关。

* <0.05

** <0.01

虽然我们证明了 SCLIWC2024 在相当数量的微博状态(一个月)中检测心理表达的有效性，但当文本长度缩短时，SCLIWC2024 是否仍然有效？为了检验 SCLIWC2024 对短文本心理表达识别的有效性，我们对一天、一周、一个月的微博状态下的 SCLIWC2024 得分(词数计分和状态数计分)和人工评分进行了 Pearson 相关分析(表 4)。结果发现从一天到一周，一周到一个月，显著相关的词类数依次减少，但都有小到中等的相关。我们的结果与 Zhao 等人(2016)的结果完全相反，这可能与人工评分者主观感受的形成机制有关，相比于一整月杂乱无序、主题不定的微博状态，评分者判断一天的心理语义表达是更容易的。而对于几乎所有 SCLIWC2024 得分与人类评分之间存在显著相关性的类别，使用 SCLIWC2024 状态数计分的方法时，相关性变得更高，这与 Zhao 等人(2016)的结果一致，表明相比词数计分，使用状态数计分能更好地利用 SCLIWC2024 词典检测心理语义表达。

表 4 不同时间跨度微博状态下人类评分与 SCLIWC2024 分数(词数/状态数)之间的相关性

	微博(月)	长文本	传统网络文本
认知历程			
allnone	**0.023/0.364****	**0.253**/0.314****	0.125/0.028
cause	**0.059/0.437****	**0.131/0.545****	0.159/0.121
insight	**0.105/0.318****	**0.044/0.592****	0.127/0.145
tentat	0.098/0.036	0.368**/0.141	0.154/0.108
情感历程			
emo_anger	**0.249**/0.291****	0.052/0.123	0.236**/0.116

续表 4

	微博(月)	长文本	传统网络文本
emo_anx	0.168/0.326**	0.098/0.148	0.144/0.274**
emo_sad	0.041/0.272**	0.325**/0.554**	0.231*/0.046
emotion	0.199*/0.253**	0.155/0.214*	0.223*/0.02
tone_neg	0.208*/0.212*	0.01/0.193*	0.11/0.012
tone_pos	0.232*/0.269**	0.200*/0.182*	0.213*/0.073
社会历程			
conflict	0.188*/0.461**	0.057/0.492**	0.019/0.204*
moral	0.181*/0.229*	0.093/0.007	0.103/0.012
polite	0.051/0.302**	0.085/0.346**	0.113/0.154
prosocial	0.109/0.360**	0.114/0.213*	0.086/0.298**
生理历程			
health	0.195*/0.455**	0.105/0.223*	0.094/0.063
生活方式			
work	0.139/0.431**	0.154/0.348**	0.205*/0.162
驱力词			
achieve	0.116/0.254**	0.115/0.252**	0.017/0.048
affiliation	0.222*/0.382**	0.036/0.324**	0.094/0.164
power	0.116/0.182*	0.149/0.255**	0.108/0.006
动机			
allure	0.103/0.312**	0.072/0.351**	0.127/0.265**
curiosity	0.012/0.026	0.112/0.014	0.178/0.022
risk	0.124/0.349**	0.293**/0.431**	0.178/0.188*
状态			
acquire	0.09/0.039	0.077/0.259**	0.056/0.029
fatigue	0.351**/0.284**	0.200*/0.326**	0.099/0.12
fulfill	0.092/0.031	0.126/0.035	0.018/0.048

注：加粗的成对数据表明，当使用状态数计分作为 SCLIWC2024 评分方法时，相关性显著且系数增大。

* <0.05

** <0.01

实验 2

在新浪微博等社交网络平台上，状态(帖子)是表达完整思想的自然单位，也是解读社交媒体用户心理表达的单位(Zhao et al., 2016)。在实验 1 中，SCLIWC2024 状态数计分在检测微博状态的心理表达方面突出，那么这种方法是否可以用来判断单个微博状态的心理意义？若能根据某一状态的心理含义，自动准确地将其归类为 SCLIWC2024 类别，则可进一步扩大 SCLIWC2024 的适用范围。为了回答这个问题，我们进行了实验 2，旨在评估 SCLIWC2024 在识别单个微博状态心理意义方面的有效性。

参考 Bantum 和 Owen(2009)以及 Zhao 等人(2016)的方法，我们使用信号检测论(Green & Swets, 1966)和信号检测指标来量化 SCLIWC2024 识别心理表达的有效性。为了估计单个微博状态中某 SCLIWC2024 类别的词是否表达了该类别的意义，我们把表达某一类别心理意义称为信号，而缺乏这种表达的则为噪声。实验采用了四个信号检测论指标：

(a) 灵敏度。被评分者识别为指示每个类别的微博状态中被 SCLIWC2024 标记为同一类别的比例；

(b) 特异性。被评分者识别为不指示每个类别的微博状态也被 SCLIWC2024 标记为与同一类别无关的比例；

(c) 阳性预测值。SCLIWC2024 标记为指示每个类别的微博状态与评分者对同一类别编码一致的概率；

(d) 阴性预测值。SCLIWC2024 标记为不指示每个类别的微博状态与评分者判断该状态与同一类别无关的相符概率。

方法

参与者和材料。对实验 1 中的 120 名微博用户，我们对每个用户随机抽取了 5 条微博状态，共计 600 条微博状态(M=36.08, SD=39.61)，作为实验 2 的样本。

人工评分。实验 1 中的 3 名评分者继续在实验 2 中进行评分，评分者被要求根据词类定义独立判断给定文本是否表达了某词类含义。评分规则与实验 1 相同，目的是识别微博状态的真实含义而不是检测任何特定的单词。在每个类别上，每个微博状态由 3 个评分者来评定，若出现不一致，则依据少数服从多数原则确定最终评定结果。我们对每对评分者在每个词类的评定进行了评分者一致性计算，kappa 得分在 0.57 至 0.93，表明不同评分者之间的一致性良好。

SCLIWC2024 评分。SCLIWC2024 的评分与实验 1 中的状态数计分相同。

结果与讨论

为了评估 SCLIWC2024 识别单个微博状态心理表达的有效性，我们使用人工评分和 SCLIWC2024 评分的数据计算了四个信号检测论指标：灵敏度、特异性、阳性预测值、阴性预测值(表 5)。

结果显示，SCLIWC2024 对心理表达识别的效度在不同指标和不同类别上表现出很大差异。除洞察词(insight)、生气词(emo_anger)、冲突词(conflict)、吸引

词(allure)、好奇心(curiosity)、获得词(acquire)、疲倦词(fatigue)外，大多数 SCLIWC2024 词类的灵敏度均高于 0.7，绝对主义思维词(allnone)、因果词(cause)的灵敏度达到了 0.95。除权力词(power)和吸引词(allure)外，大多数 SCLIWC2024 词类的特异性均高于 0.7，生气词(emo_anger)、焦虑词(emo_anx)、悲伤词(emo_sad)、冲突词(conflict)、好奇心(curiosity)、疲倦词(fatigue)的特异性达到了 0.95。阳性预测值词类之间存在比较大的差异(0.11-0.81)，并且大部分词类的值都在 0.7 以下，只有情绪词(emotion)(0.77)和正向感情(tone_pos)(0.81)的该指标值较高，即被 SCLIWC2024 识别为直接表达情绪和表达正向感情的文本中，分别有 77%和 81%的文本也被人工评分者认为直接表达了情绪或表达了正向感情。对于阴性预测值，除正向感情(tone_pos)(0.64)外，大部分词类的该指标值都非常高(0.81-1.00)，即有 36%的文本虽然没有出现正向感情的词，但语气或内容其实表达了正向情感。

表 5 SCLIWC2024 在单个微博状态评分中的平均敏感性、特异性、阳性预测值和阴性预测值

	灵敏度	特异性	阳性预测值	阴性预测值
认知历程				
allnone	0.95(0.05)	0.79(0.01)	0.36(0.03)	0.99(0.01)
cause	0.97(0.06)	0.87(0.00)	0.33(0.03)	1.00(0.00)
insight	0.61(0.12)	0.82(0.03)	0.61(0.04)	0.81(0.10)
tentat	0.78(0.04)	0.83(0.01)	0.34(0.06)	0.97(0.01)
情感历程				
emo_anger	0.60(0.11)	0.96(0.00)	0.27(0.03)	0.99(0.01)
emo_anx	0.72(0.25)	0.98(0.00)	0.17(0.00)	1.00(0.00)
emo_sad	0.82(0.13)	0.97(0.00)	0.44(0.05)	0.99(0.00)
emotion	0.70(0.09)	0.87(0.03)	0.77(0.07)	0.82(0.09)
tone_neg	0.82(0.07)	0.84(0.01)	0.51(0.04)	0.96(0.02)
tone_pos	0.76(0.04)	0.71(0.04)	0.81(0.06)	0.64(0.10)
社会历程				
conflict	0.63(0.07)	0.99(0.00)	0.62(0.03)	0.99(0.00)
moral	0.75(0.09)	0.77(0.00)	0.12(0.01)	0.99(0.01)
polite	0.78(0.02)	0.85(0.03)	0.39(0.14)	0.97(0.01)
prosocial	0.75(0.07)	0.89(0.01)	0.50(0.05)	0.96(0.02)
生理历程				
health	0.82(0.08)	0.87(0.00)	0.21(0.01)	0.99(0.00)
生活方式				
work	0.92(0.04)	0.71(0.00)	0.44(0.02)	0.97(0.02)
驱力词				
achieve	0.92(0.00)	0.75(0.00)	0.19(0.01)	0.99(0.00)
affiliation	0.78(0.08)	0.73(0.01)	0.42(0.05)	0.92(0.04)
power	0.92(0.04)	0.67(0.00)	0.11(0.02)	0.99(0.00)

续表 5

	灵敏度	特异性	阳性预测值	阴性预测值
动机				
allure	0.64(0.03)	0.62(0.02)	0.26(0.08)	0.89(0.03)
curiosity	0.42(0.13)	0.97(0.00)	0.15(0.05)	0.99(0.00)
risk	0.76(0.20)	0.93(0.00)	0.30(0.04)	0.99(0.01)
状态				
acquire	0.64(0.10)	0.83(0.01)	0.37(0.06)	0.93(0.04)
fatigue	0.39(0.10)	0.99(0.00)	0.17(0.06)	1.00(0.00)
fulfill	0.77(0.02)	0.93(0.00)	0.20(0.02)	0.99(0.00)

讨论

几十年来，语言心理学研究一直建立在“单词代表注意力”的原则上，在这方面，以 LIWC 为代表的自然语言处理工具一直是解释和语境化语言心理学研究结果的直观和可靠的框架(Boyd & Schwartz, 2021)。在这套工具的帮助下，语言心理学研究跨越了许多不同的主题，尤其是随着 LIWC-22 的面世，许多新增词类在检测文本心理语义方面表现出了很好的侦测率和有效性(Boyd et al., 2022)。因此，为了让中文化的词典更适合 LIWC-22 的框架，本文通过两项研究对 SCLIWC2024 进行了更新和验证。

在研究一，我们通过添加新增词、融合 LIWC2022 词典、大语言模型扩充词典、陈旧词处理和词类结构检查四个阶段完成了对 SCLIWC2024 的修订，得到了包含 27 个语言维度词类、41 个心理过程词类、48 个扩展词类的 SCLIWC2024 词典。在研究二，我们通过两项实验对 SCLIWC2024 词典进行了效度检验，回答了两个问题。

SCLIWC2024 检测不同网络文本心理表达的有效性。两项实验的结果均表明 SCLIWC2024 对微博短文本、长文本和传统网络文本的若干重要的心理表达的检测是有效的。实验 1 中，我们通过对 SCLIWC2024 得分和人工评分进行相关分析，在大部分词典类别中发现了显著相关性，而且在使用 SCLIWC2024 状态数计分方法时，相关系数可能更高，这表明了我们更新和扩充的 SCLIWC2024 词典对检测网络文本的心理表达是有效的。实验 2 基于信号检测论的评估显示，SCLIWC2024 词典的大部分词类都具有较高的灵敏度、特异性和阴性预测值，四项指标值均高于 Zhao 等人(2016)之前的结果。三种网络文本类型在内容和风格上的差异也反映在我们的结果中，包括单词百分比以及 SCLIWC2024 得分与人

工评分之间的相关性。单词百分比的差异反映了不同网络文本的性质，而相关系数的差异则表明了 SCLIWC2024 在不同类型网络文本中检测心理表达的有效性存在差异。

如何更有效地使用 SCLIWC2024 来检测社交网络平台短文本的心理表达。我们从文本量和评分方法两个方面对该问题进行了探讨。我们发现无论是短到一天还是长到一个月的微博状态量，SCLIWC2024 都可以有效地检测多个类别的心理语义，但随着状态数量的减少(从一个月减少到一天)，SCLIWC2024 可以有效地检测更多类别。这可能与人工评分者主观感受的形成机制有关，相比于一天，一个月的微博状态更加杂乱无序、主题不定，评分者不太容易准确把控所有文本整体表达心理语义的程度，而这恰恰表明了 SCLIWC2024 词典在处理大量短文本时，比人工评定更有优势。此外，除了使用 SCLIWC2024 词数计分(单词百分比)，还可以考虑状态数计分。在我们的结果中，相比词数计分，状态数计分通常与人类评分有更高更显著的相关性，这表明状态数计分是使用 SCLIWC2024 检测社交网络平台文本心理表达的非常有效的方法。

SCLIWC2024 对单个微博状态心理表达识别的有效性。虽然在一组微博短文本中 SCLIWC 状态数计分与人类评分之间存在显著相关性，但在单个微博状态中 SCLIWC 状态数计分识别心理表达的有效性并不是那么好。一个主要问题是阳性预测值较差，这与 Zhao 等人(2016)的结果一致。例如，生气词(emo_anger)的阳性预测值为 0.27，这表明被 SCLIWC2024 识别为直接表达生气情绪的文本中，只有 27%的文本也被人工评分者认为直接表达了生气情绪。也就是说，SCLIWC2024 对单个微博状态心理语义的判断包含较多虚报。正如 Boyd 和 Schwartz(2021)指出的那样，这是基于“单词代表注意力”原则的工具的固有缺陷，它们忽略了上下文语境。然而，考虑到 SCLIWC2024 大部分词类在灵敏度、特异性和阴性预测值上很好的表现，我们可以得出结论，即通过 SCLIWC2024 可以覆盖大部分表达所选词类含义的微博状态，但 SCLIWC2024 识别单个微博状态的心理意义的效度并不理想。

总结

本研究对 SCLIWC2024 词典进行了系统性、全面性地更新，并为其作为分

析网络文本心理语义的有效性提供了证据。我们的研究推进了 LIWC 工具中文化的进程，揭示了 LIWC 工具在处理不同主题文本时的普遍问题。

然而，LIWC 系列工具只是简单基于“单词代表注意力”的原则，虽然扩展了一个描述各种心理过程和注意模式之间相关性的网络，但它限制了使用单词计数方法探索经验问题的类型，也没有被设计用来适应语言行为的全部复杂性和功能(Boyd & Schwartz, 2021)。因此，我们仍需要思考如何使用和改进这些工具，以更好地服务于网络数据挖掘和语言心理学研究。

参考文献

岑玉珍 (主编). (2013). 汉语副词词典. 北京大学出版社.
方清明 (主编). (2017). 现代汉语介词用法词典. 商务印书馆.
黃金蘭、Cindy K. Chung、Natalie Hui、林以正、謝亦泰、Ben C. P. Lam、程威銓、Michael H. Bond、James W. Pennebaker (2012). 中文版「語文探索與字詞計算」詞典之建立. *中華心理學刊，54(2)*，185-201。 https://doi.org/10.6129%2fCJP.2012.5402.04
林瑋芳、黃金蘭、林以正、李嘉玲、James W. Pennebaker (2020). 語言探索與字詞計算詞典 2015 中文版之修訂. *調查研究-方法與應用，(45)*，73-118。 https://www.airitilibrary.com/Article/Detail?DocID=10281649-202010-202011030012-202011030012-73-118
刘月华, 潘文娱, 故韡. (2019). *实用现代汉语语法（第三版）*. 商务印书馆.
齐沪扬 (主编). (2011). 现代汉语语气成分用法词典. 商务印书馆.
中国社会科学院语言研究所词典编辑室 (主编). (2016). 现代汉语词典（第 7 版）. 商务印书馆.
Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment*, *21*(1), 79–88. https://doi.org/10.1037/a0014643
Biggiogera, J., Boateng, G., Hilpert, P., Vowels, M., Bodenmann, G., Neysari, M., Nussbeck, F., & Kowatsch, T. (2021). BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples’ Conflict Interactions. *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 385–389. https://doi.org/10.1145/3461615.3485423
Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *The University of Texas at Austin.* https://www.liwc.app
Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017
Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028
Brockmeyer, T., Zimmermann, J., Kulessa, D., Hautzinger, M., Bents, H., Friederich, H.-C., Herzog, W., & Backenstrass, M. (2015). Me, myself, and I: Self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01564
Caton, S., Hall, M., & Weinhardt, C. (2015). How do politicians use Facebook? An applied Social Observatory. *Big Data & Society*, *2*(2), 2053951715612822. https://doi.org/10.1177/2053951715612822

Che, W., Li, Z., & Liu, T. (2010). LTP: A Chinese Language Technology Platform. In Y. Liu & T. Liu (Eds.), *Coling 2010: Demonstrations* (pp. 13–16). Coling 2010 Organizing Committee. https://aclanthology.org/C10-3004

Cherrstrom, C. A. (2024). Drives and Motives During Online Degree Completion: Commonalities Among and Differences Between Hispanic and Non-Hispanic Adult Students. *International Journal of Adult Education and Technology (IJAET)*, *15*(1), 1–14. https://doi.org/10.4018/IJAET.335085

Cherrstrom, C. A., Boden, C. J., & Sherron, T. (2023). Nontraditional Students and Credit for Prior Learning—Analytical Thinking, Clout, Drives, and Motives. *The Journal of Continuing Higher Education*, *0*(0), 1–16. https://doi.org/10.1080/07377363.2023.2222247

Date, S., Deshmukh, S. N., Boyd, R., Ashokkumar, A., & Pennebaker, J. W. (2024). Designing of a Novel Framework for Marathi Natural Language Processing: MR-LIWC2015. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(11s), Article 11s.

Gao, R., Hao, B., Li, H., Gao, Y., & Zhu, T. (2013). Developing Simplified Chinese Psychological Linguistic Analysis Dictionary for Microblog. In K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, & N. Zhong (Eds.), *Brain and Health Informatics* (pp. 359–368). Springer International Publishing. https://doi.org/10.1007/978-3-319-02753-1_36

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior* (pp. xxi, 317). Univ of California Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969-2012). New York: Wiley.

Hall M., & Caton S. (2017). Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook. *PLOS ONE*, *12*(9), e0184417. https://doi.org/10.1371/journal.pone.0184417

Haupt, M. R., Chiu, M., Chang, J., Li, Z., Cuomo, R., & Mackey, T. K. (2023). Detecting nuance in conspiracy discourse: Advancing methods in infodemiology and communication science with machine learning and qualitative content coding. *PLOS ONE*, *18*(12), e0295414. https://doi.org/10.1371/journal.pone.0295414

Jordan, K. N., Pennebaker, J. W., & Ehrig, C. (2018). The 2016 U.S. Presidential Candidates and How People Tweeted About Them. *Sage Open*, *8*(3), 2158244018791218. https://doi.org/10.1177/2158244018791218

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. https://doi.org/10.1177/0261927X13502654

Konopasky, A., Durning, S. J., Artino, A. R., Ramani, D., & Battista, A. (2020). The Linguistic Effects of Context Specificity: Exploring Affect, Cognitive Processing, and Agency in Physicians' Think-Aloud Reflections. *Diagnosis*, *7*(3), 273–280. https://doi.org/10.1515/dx-2019-0103

Li L., Li A., Hao B., Guan Z., & Zhu T. (2014). Predicting Active Users' Personality Based on Micro-Blogging Behaviors. *PLOS ONE*, *9*(1), e84997. https://doi.org/10.1371/journal.pone.0084997

Lyu, S., Ren, X., Du, Y., & Zhao, N. (2023). Detecting depression of Chinese microblog users via text analysis: Combining Linguistic Inquiry Word Count (LIWC) with culture and suicide related lexicons. *Frontiers in Psychiatry*, *14*. https://doi.org/10.3389/fpsyt.2023.1121583

Markowitz, D. M. (2023). Analytic thinking as revealed by function words: What does language really measure? *Applied Cognitive Psychology*, *37*(3), 643–650. https://doi.org/10.1002/acp.4057

Markowitz, D. M., & Griffin, D. J. (2020). When context matters: How false, truthful, and genre-related communication styles are revealed in language. *Psychology, Crime & Law*, *26*(3), 287–310. https://doi.org/10.1080/1068316X.2019.1652751

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). "LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015. *PsyArXiv*, (a).

O'Dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L., & Christensen, H. (2017). A Linguistic Analysis of Suicide-Related Twitter Posts. *Crisis*, *38*(5), 319–329. https://doi.org/10.1027/0227-5910/a000443

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, *9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, Emotional, and Language Processes in Disclosure. *Cognition and Emotion*, *10*(6), 601–626. https://doi.org/10.1080/026999396380079

Samson, D. R., Clerget, A., Abbas, N., Senese, J., Sarma, M. S., Lew-Levy, S., Mabulla, I. A., Mabulla, A. Z. P., Miegakanda, V., Borghese, F., Henckaerts, P., Schwartz, S., Sterpenich, V., Gettler, L. T., Boyette, A., Crittenden, A. N., & Perogamvros, L. (2023). Evidence for an emotional adaptive function of dreams: A cross-cultural study. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-43319-z

Song Y., Shi S., Li J., & Zhang H. (2018). *Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings*. 175–180. https://doi.org/10.18653/v1/N18-2028

Stanton, A. M., Meston, C. M., & Boyd, R. L. (2017). Sexual Self-Schemas in the Real World: Investigating the Ecological Validity of Language-Based Markers of Childhood Sexual Abuse. *Cyberpsychology, Behavior, and Social Networking*, *20*(6), 382–388. https://doi.org/10.1089/cyber.2016.0657

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis* (p. 651). M.I.T. Press.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech.

Journal of Personality and Social Psychology, *118*(2), 364–387. https://doi.org/10.1037/pspp0000244

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Ulrichová, M., & Havigerová, J. M. (2023). *Medical Students' Self-Reflection of Stress Management: A Psycholinguistic Analysis*. 55–65. https://doi.org/10.15405/epiceepsy.23124.5

Varma, P., Burge, M., Meaklim, H., Junge, M., & Jackson, M. L. (2021). Poor Sleep Quality and Its Relationship with Individual Characteristics, Personal Experiences and Mental Health during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *18*(11), Article 11. https://doi.org/10.3390/ijerph18116030

Yahya, N. H., & Abdul Rahim, H. (2023). Linguistic markers of depression: Insights from english-language tweets before and during the COVID-19 pandemic. *Language and Health*, *1*(2), 36–50. https://doi.org/10.1016/j.laheal.2023.10.001

Zhao, N., Jiao, D., Bai, S., & Zhu, T. (2016). Evaluating the Validity of Simplified Chinese Version of LIWC in Detecting Psychological Expressions in Short Texts on Social Network Services. *PLOS ONE*, *11*(6), e0157947. https://doi.org/10.1371/journal.pone.0157947

Ziemer, K. S., & Korkmaz, G. (2017). Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis. *Computers in Human Behavior*, *76*, 122–127. https://doi.org/10.1016/j.chb.2017.06.038